

**DOKUZ EYLÜL UNIVERSITY**  
**GRADUATE SCHOOL OF NATURAL AND APPLIED SCIENCES**

**TEXT ANALYTICS IN STOCK MARKET**  
**PRICE PREDICTION**

by  
**Emre KARAŞAHİN**

**September, 2022**  
**İZMİR**

# **TEXT ANALYTICS IN STOCK MARKET PRICE PREDICTION**

**A Thesis Submitted to the  
Graduate School of Natural and Applied Sciences of Dokuz Eylül University  
In Partial Fulfillment of the Requirements for the Degree of Master of Science  
in Computer Engineering, Computer Engineering Program**

**by  
Emre KARAŞAHİN**

**September, 2022**

**İZMİR**

## M.Sc THESIS EXAMINATION RESULT FORM

We have read the thesis entitled “**TEXT ANALYTICS IN STOCK MARKET PRICE PREDICTION**” completed by **EMRE KARAŞAHİN** under supervision of **ASSOC. PROF. DR. SEMİH UTKU** and **ASST. PROF. DR. OKAN ÖZTÜRKMENOĞLU** and we certify that in our opinion it is fully adequate, in scope and in quality, as a thesis for the degree of Master of Science.

.....  
Assoc. Prof. Dr. Semih UTKU

Supervisor

.....  
Asst. Prof. Dr. Okan ÖZTÜRKMENOĞLU

Co-Supervisor

.....  
Prof. Dr. Emel KURUOĞLU KANDEMİR

Jury Member

.....  
Asst. Prof. Dr. Yunus Doğan

Jury Member

.....  
Asst. Prof. Dr. Mansur Alp TOÇOĞLU

Jury Member

.....  
Prof. Dr. Okan FISTIKOĞLU

Director

Graduate School of Natural and Applied Sciences

## ACKNOWLEDGMENTS

I'm extremely grateful to my thesis advisor, Assoc. Prof. Dr. Semih UTKU, for his help and support. Also, I could not have undertaken this journey without Asst. Prof. Dr. Okan ÖZTÜRKMENOĞLU, my second advisor, who was a great mentor and advisor. I would like to express my deepest gratitude to my former advisor, Prof. Dr. Adil ALPKOÇAK, who was very motivating and supported me during the whole process in the master program.

Besides, words cannot express my gratitude to my one and only wife, Melis KARAŞAHİN, who supported, motivated, and encouraged me during my thesis and never lost her faith in me.

I am also thankful to Batuhan AVLAYAN, a good friend, for his help.

Lastly, I would be remiss in not mentioning my family, especially my mother and my sister. Their belief in me has kept my self-confidence and motivation high during this process.

Emre KARAŞAHİN

# TEXT ANALYTICS IN STOCK MARKET PRICE PREDICTION

## ABSTRACT

Trying to predict the future using social media data and analytics is very popular today. With this motivation, we aimed to make stock market predictions by creating different analysis models for 10 different banks traded in “Borsa Istanbul 100” over three different groups that we selected on social media. The groups determined within the scope of the study can be detailed as tweets posted by banks from their accounts, tweets posted with the name of the bank, and tweets with the name of the bank posted from approved accounts. In our analysis, we used various variations, including the tweets' sentiments, replies, retweets, and like counts of the tweets, the effects of daily currency (Dollar, Euro, and Gold) prices, and the changes in stock changes up to 3 days. We applied some pre-processing techniques to the collected data and defined sentiment classes for sentiment analysis, created six different models, and analyzed them using 7 different classification algorithms such as Multi-Layer Perceptron, Random Forest, and deep learning algorithm. We labeled our dataset with 3 different classes to predict the stock market prices of the selected data group. According to these classes, the stock price can be positive, negative, and neutral. After all the models and analysis, we got a total of 1440 different results. According to our results, the accuracy rates vary according to the data groups and models we have chosen. The tweet group in which the name of the banks is mentioned can be shown as the most successful data group and we can easily say that there is a certain relation between social media and stock market prices.

**Keywords:** Stock market, classification, deep learning, social media, sentiment analysis.

# BORSA FİYAT TAHMİNLEMEDE METİN ANALİTİĞİ

## ÖZ

Günümüzde sosyal medya verilerini ve analitiğini kullanarak geleceği tahmin etmeye çalışmak oldukça popülerdir. Bu motivasyonla, bu çalışmada, sosyal medya üzerinden yarattığımız 3 farklı veri grubu üzerinden “Borsa İstanbul 100” içerisinde işlem gören 10 farklı banka için farklı analiz modelleri oluşturarak borsa tahmini yapmayı hedefledik. Çalışma kapsamında belirlenen veri grupları, bankaların hesaplarından atılan tweetler, bankaların adını içeren tweetler ve onaylı hesaplardan bankaların adının yer aldığı tweetler olarak detaylandırılabilir. Analizimizde tweet'lerin duygu, tweet'lere gelen yanıtlar, tweet'lerin retweet ve beğeni sayıları, günlük döviz (Dolar, Euro ve Altın) fiyatlarının etkileri ve seçilen bankaların 3 gün öncesine kadar borsa değişimleri gibi çeşitli varyasyonlar kullandık. Toplanan verilere bazı ön işleme teknikleri uyguladık ve duygu analizi için duygu sınıfları tanımladık, 6 farklı model oluşturduk. Multi-Layer Perceptron, Random Forest ve derin öğrenme algoritması gibi 7 farklı sınıflandırma algoritması kullanarak elde ettiğimiz verileri analiz ettik. Seçmiş olduğumuz veri grubunun borsa fiyatlarını tahmin edebilmek için veri setimizi 3 farklı sınıfla etiketledik. Bu sınıflara göre borsa fiyat tahminleri pozitif, negatif ve nötr olarak adlandırılabilir. Tüm modeller ve analizlerden sonra toplam 1440 farklı sonuç elde ettik. Elde ettiğimiz sonuçlara göre doğruluk oranlarının seçtiğimiz veri grupları ve modellere göre değişiklik gösterdiğini gözlemledik. Ayrıca bankaların adını içeren tüm tweet grubu en başarılı veri grubu olarak gösterilebilir ve sosyal medya ile borsa fiyatları arasında belli bir ilişki olduğunu rahatlıkla söyleyebiliriz.

**Anahtar kelimeler:** Borsa, sınıflandırma, derin öğrenme, sosyal medya, duygu analizi.

## CONTENTS

	Page
M.Sc THESIS EXAMINATION RESULT FORM .....	ii
ACKNOWLEDGMENTS .....	iii
ABSTRACT .....	iv
ÖZ .....	v
LIST OF FIGURES .....	viii
LIST OF TABLES .....	ix
<b>CHAPTER 1 -INTRODUCTION .....</b>	<b>1</b>
1.1 Background Information.....	1
1.2 Problem Definition .....	2
1.3 Aim of the Thesis.....	2
1.4 Thesis Outline .....	3
<b>CHAPTER 2 -LITERATURE REVIEW .....</b>	<b>4</b>
<b>CHAPTER 3 -DATA GATHERING .....</b>	<b>18</b>
3.1 Data Collection .....	19
3.1.1 Twitter Data .....	19
3.1.2 Stock Market Data.....	23
3.1.3 Currency and Gold Price Data .....	24
<b>CHAPTER 4 -METHODOLOGY .....</b>	<b>25</b>
4.1 Data Preprocessing .....	25
4.2 Classification and Deep Learning.....	29
<b>CHAPTER 5 -RESULTS .....</b>	<b>33</b>
5.1 Model 1 .....	35

5.2 Model 2 .....	36
5.3 Model 3 .....	38
5.4 Model 4 .....	40
5.5 Model 5 .....	42
5.6 Model 6 .....	43
 <b>CHAPTER 6 -CONCLUSION .....</b>	<b>46</b>
 <b>REFERENCES .....</b>	<b>47</b>
 <b>APPENDICES.....</b>	<b>51</b>





## LIST OF FIGURES

	Page
Figure 2.1 Parameterization values for Social Media Data .....	10
Figure 2.2 Scaling Social Media Providers with some parameters .....	10
Figure 2.3 Classification results of Chen et. al. (2019) .....	14
Figure 2.4 Proposed structure of Corosia et. al. (2019).....	16
Figure 3.1 Data related transaction flow of this paper.....	18
Figure 3.2 Twitter Post API limitations.....	20
Figure 3.3 Twitter Get API limitations.....	20
Figure 3.4 Tweet distribution between selected banks and groups .....	22
Figure 3.5 An example of created stock market data item .....	23
Figure 3.6 An example of created currency and gold price data item .....	24
Figure 5.1 Accuracy of the Halkbank contained for the Model 6 .....	33
Figure 5.2 Accuracy of the GarantiBBVA contained for the Model 5.....	34
Figure 5.3 Accuracy of the Vakıfbank contained for the Model 6.....	34

## LIST OF TABLES

	Page
Table 3.1 Twitter account and stock market names of the selected banks .....	18
Table 3.2 Twitter API limitations .....	19
Table 3.3 Example Twint commands for getting data.....	21
Table 3.4 JSON file names and included Tweet counts .....	22
Table 4.1 Some examples of Tweets before and after preprocessing.....	26
Table 4.2 Some examples of sentiments for Tweets .....	28
Table 4.3 Summary of created models with details.....	31
Table 5.1 Model 1 classification results .....	35
Table 5.2 Model 2 classification results .....	37
Table 5.3 Model 3 classification results .....	38
Table 5.4 Model 4 classification results .....	40
Table 5.5 Model 5 classification results .....	42
Table 5.6 Model 6 classification results .....	43

# CHAPTER 1

## INTRODUCTION

### 1.1 Background Information

Social media specifies websites and applications which are planned to permit people to share content rapidly, effortlessly, and in real-time. It is used by everyone from 7 to 70 and contains very large open-ended data, which can be called big data. Users can share their opinions with other users on social media platforms, even some users lead the market with comments they have made. For example, some users review a product and share their experience with their followers, and they present their positive or negative opinions about this product to their followers. These opinions can affect the sales of the product in a very significant way. In this case, it clearly shows us the effects of the views of people on social media. For this reason, some companies make agreements with some social media accounts which have high followers to advertise their products. Also, users share positive or negative opinions about a company or a product through social media. These shares can greatly affect the brand values of companies. For this reason, most firms have opened accounts on social media to support their customers.

A stock market is the accumulation of purchasers and vendors of stocks, which define ownership requisitions on enterprises. These may contain securities registered on a common stock exchange, besides stock that is only traded confidentially. Stock market estimation is the action of seeking to decide the oncoming worth of firm stock or another financial portfolio traded on an exchange. The successful estimation of a stock's oncoming worth could yield considerable gain.

An exchange rate is a notion that states the value or number of foreign currencies in terms of foreign changes. For instance, the worth of the dollar in Turkish lira is shown with the dollar/TL rate. Likewise, its worth in Euro is shown with the Euro/TL rate. Currency is essentially no different from the products you see in the market. The

value of a product you see in the market is determined by the laws of supply and demand. The same is true for currencies. The value of 100 TL depends on the demand for 100 TL in the market. If people demand Turkish Lira, the value of the Turkish Lira increases.

## **1.2 Problem Definition**

Today, social media can be considered as an endless source of information and content. Making various interpretations using the data here can lead researchers to very important and interesting results.

In this study, we wanted to examine how effective the shares made on social media are in a certain area and to examine whether the shares made have an effect on concrete events.

With this motivation, we tried to make stock market predictions by using various text analytics techniques by addressing the banks that are traded on Borsa Istanbul 100 and their posts on social media.

## **1.3 Aim of the Thesis**

In this study, we will try to make estimations about the market values of the companies operating in the stock exchange by using some elements which are the posts shared on social media accounts, the number of likes of these posts, the comments on these posts, the sharing number of these posts. We want to observe whether there exists any association between social media usage, the company's stock price, and currency prices. For this, we will collect daily data from desired social media accounts that are related to the stock market or trading. After that, we will analyze and visualize the collected data and make predictions with it.

## 1.4 Thesis Outline

The thesis consists of six main chapters. The rest of this paper is organized as follows. In Section 2, we review previous related works. In Section 3, we explain how to collect data and prepare a dataset. In Section 4, we describe the methodology and used techniques in this study. In Section 5, we review solutions and results. In Section 6, we review the conclusion of the study.



## CHAPTER 2

### LITERATURE REVIEW

Twitter is one of the most popular microblogging and social networking platform where people can freely share their thoughts and ideas. It allows users to talk and discuss certain topics which are called hashtags. The stock market is one of the most talked-about issues on social media. Predicting pricing on the stock market is still one of the most challenging problems. In this study, we aim to find out whether conversations on social media have an influence on pricing in the stock market.

In this section, we chronologically reviewed the studies and the articles which are related to the stock market and social media.

Blankespoor et al. (2014) express the opinion that the statements made by companies, especially by small-scale companies, on their announcement platforms do not have a very high effect on the movement in the stock market, since they do not spread to large masses. For this reason, in their studies, they examined the consequences that would occur if these announcements were spread to a wider audience with direct-access information technologies like Facebook, Twitter, RSS, etc. The team, which examines the announcements made on Twitter as a social platform, analyzes the differences between the announcements made by technology companies on social media and the announcements they made with the press releases. They collected all the tweets belonging to 102 Twitter accounts, with a tool that is written in the PERL programming language. Also, they included 233 press releases from 79 companies in the dataset. The team defines two arguments to defend their thoughts. The first is the impact of press releases on social media on stock prices. They compare their results with the effect of the announcement made on the company's site. According to the findings, they observe that the statements made on social media have a positive effect. Secondly, they examine the effect of the trade dimension on transactions and examine the relationship between retail investors and institutional investors. The team finds poor proof that small traders are trading more thanks to social media. As a result of this study, their method ensures a unique path to isolate the impact

of broader spreading like social media. This adds to the disclosure literature by showing that, in addition to the impact of the information in the description, the wider dissemination of this information can have real market consequences.

Nassirtoussi et al. (2014) provide a systematic review of text mining for market prediction. According to the team, there is no versatile theoretical and technical framework for the text mining approach for market prediction in the best way. They said that the lack of clarity on the subject is due to its interdisciplinary nature. They provide an interdisciplinary nature and contribute to the creation of a clear framework for discussion. They examine studies of online text mining-based market prediction and prepare a list of the common components for all studies. In addition, they compare each system with the others and identify their main differentiating factors. In summary, they have contributed to the literature in four different ways.

- The basic concepts of economics and information sciences were examined by considering the determined research topic. The relationship of these concepts with the proposed solutions is explained.
- The most remarkable literature study in the past has been examined given today's conditions.
- The main differentiating factors among available studies were identified and used to compare and contrast existing solutions.
- The subjects that can be used in future studies and which are less used in the literature were determined.

Liu et al. (2015) suggest a new pattern for both defining homogeneous stock sets and predicting stock change based on company-specific social media metrics. As stated in the study, companies are now able to communicate with people more easily and quickly thanks to social media. The team suggests that this communication easiness can be used to keep track of how companies change overstocks. Their empirical results show that a company that is publicly traded and has an official Twitter account is more active on the stock market than companies without an official Twitter account. They defined the number of followers, the ratio of the number of followers to

the number of followers, and the ratio of the number of followers to the number of tweets as microblogging metrics. They also add firm follower's metrics into these metrics. Because some firms may have some fake accounts as a follower. They also introduced a control metric which is called market capitalization. They used the examples of firms listed on two stock exchanges which are the NYSE and NASDAQ to explore the effect of suggested social media metrics on defining homogeneous stock sets. They get the list of public companies in the two national markets from BvD's database which is called OSIRIS and then they set the official Twitter accounts. They defined 293 American firms with official Twitter accounts and product support, or department accounts are also excluded and wrote a web crawler for collecting data from Twitter and getting five days of data. Financial data was collected from the CRSP database, and it includes data from the beginning of 2013, to the end of 2013. According to their findings, these metrics not only predict the movement of stocks but also significantly increase the accuracy of investment estimates compared to industry categories.

Lee et al. (2015) examine how social media disclosure affects the capital market in case of product recalls. The team says that social media allows companies to announce to large masses directly and quickly compared to traditional announcement channels. They used 405 product recalls from 2000 to 2012 to find out any relation between social media effects and product recall announcements. This product recall data is collected from the CPSP website. The team collects data from Twitter, Facebook, RSS, and corporate blogs. They see that social media posts generally lighten the negative price response to product recall announcements. They find that the companies which are recalling their products through any of the four social media platforms have a less negative price response than firms without social media accounts. They also show that announcements from social media in the context of a product recall affect company market value.

Nguyen et. al. (2015) aims to create a model that predicts price movements (up or down) by using sentiments from social media data. The point that distinguishes this paper from other studies in the literature is they involved the emotional analysis of



certain company issues rather than general mood analysis in their model. They used two datasets while creating that model. One of them contains historical prices, and the other one is a message board dataset. They collected adjusted close prices for the historical prices dataset from Yahoo Finance for eight-teen companies from July 23, 2012, to July 19, 2013. For the message board dataset, they collect all tweets, including the main post and its replies. The tag sentiments as Strong Buy, Buy, Hold, Sell, and Strong Sell. They compare their method with the historical price method and human sensitivity method. They compared accuracy over 18 stocks in a one-year transaction. According to their results, they achieved 2.07 better accuracy than the model which is using a historical price method. They also tested stocks that are difficult to predict, and their model has 9.83% better accuracy than the historical price method and 3.03% better accuracy than the human sensitivity process.

Akmese et al. (2016) aim to evaluate and analyze the relationship between the efficient use of social media and financial performance like sales, profits, market values, etc. In addition, the researchers aimed to find the differences between the financial performances of companies that are traded on Borsa Istanbul (BIST) and those that have and do not have social media accounts in the tourism sector. They selected 11 tourism companies traded on BIST and limited their working range to 2014. The team created the financial situation dataset with the data collected through the Public Disclosure Platform website<sup>1</sup> and created the stock market dataset of the companies selected from the tourism sector with the data collected from the Borsa Istanbul website<sup>2</sup>. In addition, the social media accounts of companies other than the selected companies and their social media activity status were examined. They applied Kolmogorov-Smirnov and Shapiro-Wilk tests before analysis to test whether the data distribution was normal, and they saw that the data distribution was not normal. So, they used the Mann-Whitney U test. According to their results;

- The average net profit of companies with social media relationships is higher than those without social media activity.

---

<sup>1</sup> <http://www.kap.gov.tr>

<sup>2</sup> <http://www.borsaistanbul.com>

- Social media does not have a significant effect on the net sales of tourism companies traded on BIST and with or without social media relations.
- Having a social media account has a relation to the market values of companies. Tourism companies that are traded on BIST and have a social media account have higher market value than the other ones which have not.

Sun et. al. (2016) aims to predict the stock market by exploring the potential uses of texts on microblog sites. They present a model which is based on text analysis to predict the stock price. They collected about forty-five million tweets between January 1, 2011, and August 31, 2015, from StockTwits. Tweet data includes text, follower and following counts, posted tweet time, etc. However, they just used text and tweet time for their analysis. They applied some text mining techniques on the R platform for preprocessing text data. Preprocessing includes URL removing, emoji removing, stop-word elimination, and converting text to its lowercase form. They prepared word count charts from preprocessed text to show the most popular text from tweets and realized that there is a negative relationship between word count and firms' prices. They collect stock data from the S & P 500 index and remove stock with low volume. In the end, they have 420 company data in their stock dataset for all trading days between their timelines. They created a sparse matrix factorized model and did daily and intraday predictions. According to their findings, they achieved better results compared with basic models. They found 51.37% accuracy for daily prediction. They also found that increasing the frequency of predictions is not affecting prediction accuracy.

López-Cabarcos et. al. (2017), examine the social network action of technical and non-technical investors, analyzing their impact on market risk and the differences between them. For tracking social network activity, they used StockTwits.com. They aimed to find out whether the type of investor made a difference in social media activity and what values from investor profiles made an impact on the market. They can analyze individual efficiency for some predictors according to a categorical variable with the usage of the logit and probit model. They used experience in making investments and the number of followers as parameters. They divided users into two

groups, users using technical and fundamental analysis. They used investors' tweets on StockTweet from 2009 to 2015 as a dataset and tweet sentiments are calculated using the sentiment analysis software of Stanford University, Stanford CoreNLP Natural Language Processing Toolkit. For financial data, they used the VIX index on the Chicago Board Options Exchange website. They collected tweet posts about 133,931 stocks regarding the S&P 500 index. According to their results;

- Social media usage has an impact on stock market activity.
- This impact can be changed according to the type of investor.
- Non-technical investors' post has an effective risk perceived in the market, but technical investors do not.

Coyne et. al. (2017), aims to predict stock prices using social media data. They used StockTweets as a social media service and created a dataset with this between May 2016 and April 2017 and searched for a large number of stocks and market capacity for better accuracy. They applied some text mining techniques for preprocessing tweets like removing stop words. They used three different models: Linear Regression, Sentiment Prediction, and Smart User Classification. With linear regression, they achieved 52.45% average accuracy. They predict sentiments under three different classes which are positive, negative, and neutral. For smart user classification, the team used the number of likes, follower count, and how often the user is correct values as a parameter. With this structure, they can identify users which post correct predictions on StockTweets and achieve 64.3% accuracy.

Kaushik et. al. (2017), designed this paper to explore in elaborately India's most traded companies which exist in social media and how this existence affects their market values. They aim to find out if there is any relation between social media usage with companies' stock prices. They used National Stock Exchange (NSE) stock prices for NIFTY 51 firms for the database, social media activities, monthly posts, and monthly responses to users were saved on four social media platforms which are Facebook, Twitter, YouTube, and LinkedIn. They parameterized values for social media data like Figure 2.1.

S. No.	Facebook	Twitter	LinkedIn	YouTube
1.	Account availability (yes/no)	Account availability (yes/no)	Account availability (yes/no)	Account availability (yes/no)
2.	Digital integration <sup>a</sup> (yes/no)	Digital integration <sup>a</sup> (yes/no)	Digital integration <sup>a</sup> (yes/no)	Digital integration <sup>a</sup> (yes/no)
3.	No. of followers	No. of followers	No. of followers and no. of employees	No. of subscribers
4.	No. of posts monthly <sup>b</sup>	No. of posts monthly <sup>b</sup>	No. of posts monthly <sup>b</sup>	No. of posts monthly <sup>b</sup>
5.	Responses to users' comments	Responses to users' comments	Responses to users' comments	Responses to users' comments

Figure 2.1 Parameterization values for Social Media Data

Also, they indexed all parameters with some scale values like Figure 2.2.

Parameter	Index	Facebook	Twitter	LinkedIn	Youtube	Total scale
Digital integration (yes/no)	Digital integration score	1	1	1	1	4
No. of followers/subscribers	SM popularity score	3	3	3	3	12
No. of posts/tweets/video uploads/monthly	SM activity score	3	3	3	3	12
User interaction (yes/no)	SM user interaction score	1	1	1	1	4

Figure 2.2 Scaling Social Media Providers with some parameters

The author used ANOVA for their analysis. According to their findings, the most developed social media score is the financial services and IT sectors, followed by the telecom, automobile, and consumer goods sectors. They stated that there is no satisfactory relationship between the social media activities of the companies and market prices.

Ruan et. al. (2018), propose to use the trust relationship between users to increase the relationship between social media and the stock market. They used real stock market data as the basis for their trust management system and to see the correlation between Twitter sentiment value and abnormal stock returns for most mentioned eight firms in the S&P 500, they collected tweet data from Twitter. On average, in their dataset, each firm had more than forty daily tweets. They collect stock from Yahoo Finance. They developed a trust management system for Twitter users and used the Pearson Correlation Test for testing their system for the trading days between 01/01/2015 and 08/31/2015 They also give some weight to Twitter accounts according to follower count for the trust management system. Unlike many existing studies that

treat all authors equally or ignore authors' identities, this study considers tweet writers in addition to analyzing the emotions of tweets. Their results show that using their proposed method, Twitter sentiment value reflects anomalous stock returns better than the other two methods which are traditional, meaning treating all authors equally importantly or weighing on follower numbers.

Wu et. al. (2018), analyzed the practicality of journals for estimating stock revenues on the Taiwanese stock market. They used text mining techniques on business journals, converting papers using a keyword matrix, and then converting the outcomes into news variables. They added news variables into their regression model with macroeconomic factors to explore the role of news variables in predicting stock market revenues. They conduct sample forecasting analysis to explore the predictive capability of statistics / dynamic models depend on root mean square error (RMSE). They collected economic data, stock prices, etc. from the Taiwan Economic Journal database and they used the Knowledge-Management Winner newspaper database for economic journals. Their datasets include data between January 2008 and December 2014. According to their findings, there are certain relations between news and the stock market. Both items (dynamic and static) forecast assessments bring to light that the addition of news variables decreases RMSE.

Chahine and Malhotra (2018), focused on examining the market reaction at the time of the creation of a Twitter platform for 312 firms from the Fortune 500 firms such as in Culnan et al (2010). For determining the effectiveness of social media platforms on company worth, the event history analysis (EHA) was used. Here, the authors designed a framework to investigate the effect of a historical phenomenon on the US Fortune 500 companies that developed a Twitter platform. According to their outcomes;

- The market reactions of companies that communicate with their followers in two ways are higher than those of companies with one-way communication.
- The market reactions are also higher for small firms and firms with family and/or controlling shareholders.

Siikanen et. al. (2018), investigates the relationship between social media data which is Facebook and investors investment decision processes in stock markets, with data on investors' transactions at Nokia. They claimed that buying and selling judgments were correlated with their social media dataset, particularly for inactive households and nonprofits. The authors used a unique investor-level shareholding registration dataset. The dataset included the mercantile of all Finnish financiers over various years. The aim of the work is how social media data connect to financiers' judgements to boost or reduce their standings. Varied financier societies, including financial foundations, nonprofit organizations, and households, and their commerce in Nokia stock addressed this question. Here, the dataset was collected from daily numbers of posts and involved comments, likes, and shares on Nokia's Facebook page from beginning of 2010 till the end of 2016 by using the Social Data Analytics Tool (SODATO). Also, the announcement data is gathered from NASDAQ OMX Nordic's website. Their results show that the buying and selling decisions of investors in different groups have an absolute relationship with social media data.

Liu et. al. (2018), explores how news media affects Chinese stock markets. They collected daily data from 360 and Baidu search engines and financial data coming from Shanghai Stock Exchange and the Shenzhen Stock Exchange and collected from Sina Weibo and Hexun finance using web crawling techniques for the period 2009 to 2016. They discover that stock trading capacity and turnover rates are positively correlated with new media activity, while stock returns are negatively correlated and also vice-versa. By comparing the influence of the traditional with the new media, they find a delayed impact in traditional media. They contributed the literature as follows;

- By reaffirming the influence of the media on new information channels, they show that the tweet can affect the financial market.
- They explore user investment behavior for different new media channels.

Chen et. al. (2018), focused on the analysis of the news content crawled from official accounts selected from Sina Weibo similar to Liu et. al. (2018) and by taking sentiment features and Latent Dirichlet Allocation (LDA) features. Also, the authors

used the features as inputs for a hybrid model called Recurrent Neural Networks (RNN)-boost to predict the stock volatility in the Chinese stock market. Their dataset consisted of two parts: the historical price of the stock market (China Shanghai Shenzhen 300 Stock Index) and the journals content on social media. They select the period of study between January 1, 2015, and February 14, 2017, to conduct the experiments which include 513 trading days. They suggest a hybrid model which is called RNN-Boost and achieved 66.54% average accuracy with this model. They contributed the literature as follows;

- To simplify the estimation of the stock index, they use journals content from online social media for analysis rather than traditional news media.
- They suggest LDA attributes to increase the efficiency of the estimation model and consolidate them with sensitivity characteristics and other technical indexes.
- They proposed a hybrid model which includes RNN and Adaboost to estimate the stock index, and trials to indicate their model performance.

Rosati et. al. (2019), want to understand whether alternative communication channels like social media decrease the cost of a data infringement, they investigate the impact of affected firms' exposure to social media on the share cost response to a data infringement notice. They selected Twitter as a social media platform and created a dataset beginning from the list of violations that happened from January 2011 through December 2014 compiled by Privacy Rights Clearinghouse. Their final dataset includes 87 cases suitable to 73 individual companies. They gathered data from three different sources. Daily stock price and market index data were collected from Thomson Reuters Datastream. They collected infringements data from journal writings using Lexis-Nexis PowerSearch and Twitter data collected from TwitterCounter. According to their findings, the possible negative effects of social media communication weigh down the possible utilities in the context of data infringement notices. Disclosure of a data infringement on social media complicates the negative cost reaction to the notice.

Chen et. al. (2019), applied seven different data mining techniques to estimate the stock price movement of the Shanghai Composite Index. They collected financial stock comments and their goal is analyze individual sentiment by calculating the accuracy of using seven machine learning algorithms to classify collected comments as two classes which are positive and negative. The algorithms contain Support Vector Machine, Logistic Regression, Naive Bayesian, K-Nearest Neighbor, Decision Tree, Random Forest, and Adaboost. They applied some text mining techniques to the data like tokenization, stop word removal, etc. Comment data is collected from the Eastmoney financial forum and Shanghai Composite Index stock price data is collected via Tushare. Their classification results are shown in Figure 2.3.

Algorithms	Accuracy	Positive			Negative		
		Precision	Recall	F1-score	Precision	Recall	F1-score
<b>LinearSVC</b>	<b>0.8816</b>	<b>0.8805</b>	<b>0.8825</b>	<b>0.8815</b>	<b>0.8824</b>	<b>0.8801</b>	<b>0.8813</b>
LogisticRegression	0.8809	0.8791	0.8832	0.881	0.8828	0.8782	0.8804
Multinomial NB	0.8796	0.8821	0.876	0.879	0.8767	0.8832	0.8799
KNN	0.8201	0.8071	0.8404	0.8234	0.8336	0.7991	0.8159
DecisionTree	0.7994	0.8169	0.7711	0.7933	0.7833	0.8272	0.8046
RandomForest	0.8137	0.8662	0.742	0.7992	0.7739	0.885	0.8256
AdaBoost	0.7719	0.7973	0.7989	0.7654	0.8253	0.7594	0.7666

Figure 2.3 Classification results of Chen et. al. (2019)

Dieijen et. al. (2020), explores whether volatility in user-generated content (UGC) will affect changes in stock revenues or not. UGC includes users' tweets, blog posts, and Google searches. The metrics of UGC are the density of positive tweets, the density of negative tweets, the density of blog posts, the volume of Google searches, etc., and the Support Vector Machine algorithm applied for sentiment classification for negative and positive. This study consists of two different studies. One of them is to use daily data which is collected from tweets and blog posts about Apple's iPhone between January 3, 2007, and March 30, 2010. The second research focalizes on the airline industry as both airlines and their clients are too operative on social media.



They use day-to-day data from first day of July 2013 to last day of June 2014 on JetBlue, Delta, United, and Southwest airlines which are collected from tweets. This study conduces to the literature as follows;

- Initial research to explore the existence of shock and volatility spreads between UGC and stock revenues.
- They are investigating whether fluctuations in expansion rates of UGC volume are affected by introducing new products or other organizational events related to the corporation.
- The first article in marketing uses the Multivariate GARCH BEKK model to examine volatility.

Tonghui et. al. (2020), proposes a new measure of microblogging data (the Sina Weibo Index) using daily publishing, interpretation, and tagging activities for the certain stock index on the Sina Weibo like Liu et. al. (2018) and Chen et. al. (2018). They prepare a dataset that contains daily relevant posts, comments, and tagging activities for the CSI 300 index on the Sina Weibo for only trading days. By comparing linear and nonlinear methods, they discover that there are dynamic and close relationships between Sina Weibo and stock market volatility from a nonlinear point of view. According to their results, they see that the trend of change in the Sina Weibo Index is highly correlated with stock market volatility.

Carosia et. al. (2020), purposes to manage research of the Brazilian stock market movement through sentiment analysis in Twitter data. Their studies include data covering the term of the 2018 presidential elections between 01 October 2018 and 31 December 2018. They developed sentiment analysis methods for the Portuguese language with a comparison of different machine learning methods. Their proposed structure is shown in Figure 2.4.

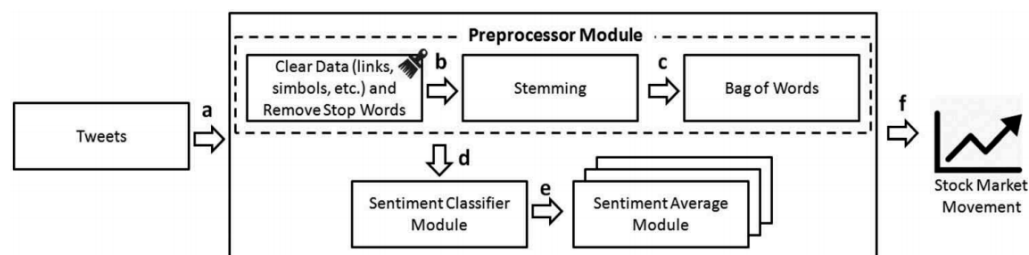


Figure 2.4 Proposed structure of Corosia et. al. (2019)

This study proves that verifying the relationship between dominant sentiment in social media and stock market movements is possible with 3 perspectives: the certain number of tweets, sentiment weighted by favorites, and sentiment weighted by the number of retweets.

Jiao et. al. (2020), investigate stock volatility and the impact of traditional news media and social media on the turnover of coverage. They use a unique panel dataset on media coverage from the Thomson Reuters MarketPsych Indices (TRMI) database. They focalize on the time range between January 2009 and December 2014. According to the results, while high social media coverage at the stock level estimates the ensuing high return volatility and trading activity, high news media coverage estimates vice-versa. This study is within the first to straight comparison news and social media and the major additive is to indicate that social media and news media have distinct correlations with stock prices.

Khan et. al. (2022), use machine learning algorithms on social media and financial news data to explore the effect of this data on stock market estimation accuracy for the 10 days. They chose Twitter as the source of social media data and used cashtags(\$) as a search query, financial headline news gathered from Business Insider, and historic data collected from Yahoo Finance. They preprocessed all text data which includes methods like tokenization, removing URLs, tags, stop word elimination, etc. Sentiment analysis of processed tweets and financial news is applied using Stanford NLP's Stanford sentiment analysis package. According to their empirical outcomes, the highest estimation accuracy is reached using social media and financial news with

80.53% and 75.16%, respectively. They find that some stock exchanges are hard to predict, some of the stocks are more affected by social media, and some of them are affected by financial news. For the random forest classifier, the highest accuracy rate was found as 83.22%.



## CHAPTER 3

### DATA GATHERING

This section describes the creation of the dataset to be used in the study, how the data for this dataset is collected and what stages it goes through. The details of the transactions performed in this section are shown in Figure 3.1.

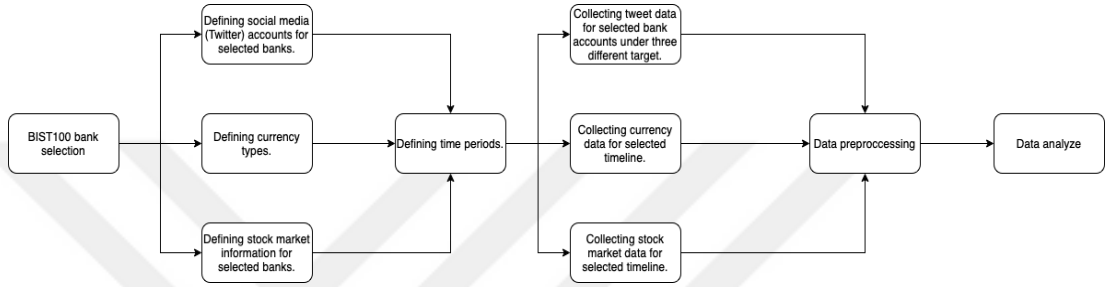


Figure 3.1 Data related transaction flow of this paper

Borsa Istanbul, also known as BIST, ensures oversight and exchange services to Turkish and foreign banks and brokerage houses running in the capital market opened in Turkey in 1985. In this study, we selected Turkey's ten leading banks operating in BIST as a data source. DenizBank was excluded from the scope as it did not trade on BIST within the specified period. The banks selected within the scope of the study are in alphabetical order: Akbank, Albaraka, Garanti BBVA, Halkbank, İş Bankası, QNB Finansbank, Şekerbank, Türkiye Sınai Kalkınma Bankası (TSKB), Vakıfbank and Yapı kredi. You can see the Twitter account and stock market name details of the selected banks in the Table 3.1.

Table 3.1 Twitter account and stock market names of the selected banks

Name	Twitter Account	Stock Market
Akbank	@Akbank	AKBNK
Albaraka	@albarakacomtr	ALBRK
Garanti BBVA	@GarantiBBVA	GARAN
Halkbank	@Halkbank	HALKB
İş Bankası	@isbankasi	ISCTR
QNB Finansbank	@qnbfinansbank	QNBFB
Şekerbank	@sekerbank	SKBNK
TSKB	@TSKB Turkey	TSKB
Vakıfbank	@VakıfBank	VAKBN
Yapı kredi	@YapiKredi	YKBNK

Along with all this data, the details of the currency (dollar and euro) and gold prices for the selected period were also collected by writing a web crawler.

### 3.1 Data Collection

In this section, how the data is obtained will be explained under three subheadings: Twitter data, stock market data, and currency data.

#### 3.1.1 Twitter Data

Twitter is a social networking and microblogging service founded in California, the USA in 2006. On Twitter, which is one of the most used social media applications, registered users can share texts, photos, news, and videos which are called “tweets”. Whether or not the tweets are visible by everyone is entirely up to the user's choice. According to “Statista” daily active user count is equal to 211 million for Twitter in whole world in the third quarter of 2021.

Twitter has some limitations which are already defined on its website. These limitations are shown in the Table 3.2.

Table 3.2 Twitter API limitations

Title	Description
Daily Direct Messages	1000 messages per day.
Tweets	2400 per day.
Changes to account e-mail	4 per hour.
Daily following	400 per day.
Account based following	Some limitations after 5000.

Twitter also has created the Twitter API and made it available for developers to use. Twitter API provides programmatic access to Twitter in different ways. Developers can access Twitter essentials like Tweets, Direct Messages, Lists, users, etc.

Since there are too many requests on the Twitter API, Twitter has set some limits to share usage. Post API limits are shown in Figure 3.2.

Endpoint	Rate limit window	Rate limit per user	Rate limit per app
POST statuses/update	3 hours*	300*	300*
POST statuses/retweet/:id	3 hours*	300*	300*
POST favorites/create	24 hours	1000	1000
POST friendships/create	24 hours	400	1000
POST direct_messages/events/new	24 hours	1000	15000

Figure 3.2 Twitter Post API limitations

Also, some of the Get API limits are shown in Figure 3.3. Detailed info can be found on the Twitter Developer website.

Endpoint	Requests / window per user	Requests / window per app
GET account/verify_credentials	75	0
GET application/rate_limit_status	180	180
GET favorites/list	75	75
GET followers/ids	15	15
GET followers/list	15	15
GET friends/ids	15	15
GET friends/list	15	15
GET friendships/show	180	15
GET geo/id/:place_id	75	0
GET help/configuration	15	15
GET help/languages	15	15
GET help/privacy	15	15
GET help/tos	15	15
GET lists/list	15	15
GET lists/members	900	75

Figure 3.3 Twitter Get API limitations

Within the scope of our work, we started to look for alternatives to obtain the data to obtain more consistent results and not be stuck with Twitter API limitations. That's why we decided to collect Tweets via Twint.

Twint is an advanced Twitter scraping instrument developed in Python that let us extracting Tweets from Twitter profiles without using Twitter's API. It allows you to get tweet data from predefined topics, desired users, trend topics, and hashtags, and supports cashtags. Twint also has some limitations which are related to user scrolling. This affects the profile and favorite section of about 3200 tweets. Twint also can show banned account history.

Twint has different storage options like writing to a file (CSV, JSON SQLite), and ElasticSearch (Graph Visualization).

We listed some of the example commands for getting data from Twint in Table 3.3.

Table 3.3 Example Twint commands for getting data

Twint Command Explanation	Command
Get all @akbank tweets since the desired date and export as a JSON file.	<code>twint -u akbank --since 2019-09-01 -o akbank.json -json --lang tr</code>
Get all tweets which are contained "akbank" keyword since desired date and export as a JSON file	<code>twint -s akbank --since 2019-09-01 -o akbank_contained.json -json --lang tr</code>
Get all tweets which are posted from verified accounts and contained "akbank" keyword since desired date and export as a JSON file	<code>twint -s akbank --verified --since 2019-09-01 -o akbank_contained_verified.json -json --lang tr</code>

We completed the relevant data collection processes for ten different banks traded in BIST100 under three different headings which are called a bank, bank name contained, and bank name contained verified. The time range of the collected data is 01/09/2019 to 25/09/2020. At the end of the tweet gathering, we have thirty different JSON files with a total file size of about 500MB.

Detailed JSON files and tweet numbers are listed in Table 3.4.

Table 3.4 JSON file names and included Tweet counts

JSON File Name	Tweet count
Akbank.json	313
Akbank_contained.json	57068
Akbank_contained_verified.json	2554
Albaraka.json	405
Albaraka_contained.json	10410
Albaraka_contained_verified.json	240
GarantiBBVA.json	478
GarantiBBVA_contained.json	65408
GarantiBBVA_contained_verified.json	3407
Halkbank.json	440
Halkbank_contained.json	113228
Halkbank_contained_verified.json	5402
Isbank.json	222
Isbank_contained.json	33600
Isbank_contained_verified.json	738
QNBFinansbank.json	212
QNBFinansbank_contained.json	9262
QNBFinansbank_contained_verified.json	546
Sekerbank.json	137
Sekerbank_contained.json	7157
Sekerbank_contained_verified.json	200
TSKB.json	346
TSKB_contained.json	35752
TSKB_contained_verified.json	104
Vakifbank.json	307
Vakifbank_contained.json	102494
Vakifbank_contained_verified.json	3531
Yapikredi.json	144
Yapikredi_contained.json	27634
Yapikredi_contained_verified.json	895

When the tweet distributions in the table are divided into their groups, the obtained percentiles are shown in the Figure 3.4. below.

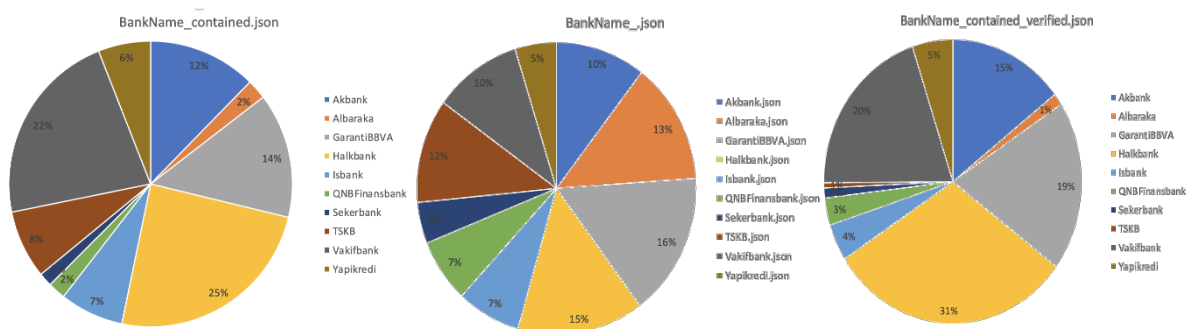


Figure 3.4 Tweet distribution between selected banks and groups



### 3.1.2 Stock Market Data

The stock market is the organized market where tangible assets such as stocks and state securities, commodities, currencies, futures, and options contracts are merchandised and proposed to the public, and where these products are purchased and sold safely. The organized stock exchange of Turkey is “Borsa İstanbul”.

BIST 100 index is the master index used to evaluate the performance of the top 100 stocks in terms of market and mercantile capacity traded in Borsa Istanbul. The operation code of the BIST 100 index, which is also acknowledged as an index of the Turkish stock market, is XU100.

BIST 100 index, one of the most popular indices of Borsa Istanbul, is an index that is carefully followed by all major investors. The biggest cause for this is the stock market's fall and rise interpretations are made by considering the BIST 100 index. In summary, the stocks of these 100 firms indicate the average actions of the stock market.

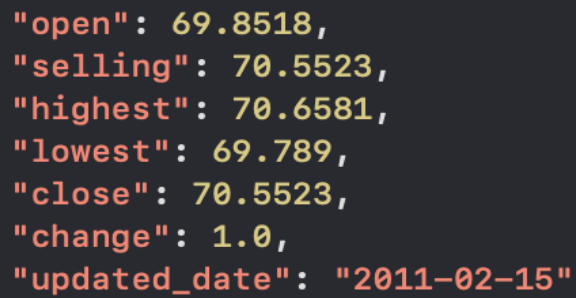
Within the scope of this study, we studied the stock market data of 10 popular banks traded in BIST 100. For this reason, we used a web scraping tool to access stock market data for the date range we selected. We collected daily stock prices, high and low prices, volume, and percentage of the daily change for every selected bank and stored them as a JSON file. As a result, we collected 365 different stock price data for every single bank. An example item for stock market data is shown in Figure 3.5.

```
"Date": "Feb 03, 2021",  
"Price": 2.19,  
"Open": 2.15,  
"High": 2.28,  
"Low": 2.15,  
"Vol": "133.91M",  
"Change": "2.82%"
```

Figure 3.5 An example of created stock market data item

### 3.1.3 Currency and Gold Price Data

Currency is expressed as a coefficient showing the value of the currency of any country against the currencies of other countries. Within the scope of our study, we created our dataset containing Dollar, Euro, and Gold prices in the date range of 10 years from February 2011 to February 2021 using the web scraping tool. We defined a common document format for all price data which contains the date, daily opening, selling, highest, lowest, and closing prices, and percentage of the daily change and store them as a JSON file. There are approximately 2600 different items in each dataset. An example item for currency and gold price data is shown in Figure 3.6.



```
"open": 69.8518,  
"selling": 70.5523,  
"highest": 70.6581,  
"lowest": 69.789,  
"close": 70.5523,  
"change": 1.0,  
"updated_date": "2011-02-15"
```

Figure 3.6 An example of created currency and gold price data item

## **CHAPTER 4**

### **METHODOLOGY**

Within the scope of the study, we aim to predict the status of stock market price changes into three classes as positive, negative, and neutral. For this purpose, we created the larger scale dataset combining the social media data and foreign exchange prices. In this chapter, we will apply the methods and talk about the experiments using our dataset under the headings of the models we have determined.

#### **4.1 Data Preprocessing**

In this section, we will talk about the methods we used to create the dataset that we will use as a source in the analysis phase. Before applying the methods, we performed the preprocessing steps with custom Python scripts. First, we took the exchange currency, gold prices, and all types of bank tweets that we mentioned in the data collection stage as raw data. We extracted non-trading day tweets from the collected data. A trading day known as non-weekends, any day that is a statutory holiday, or any day that banking corporations need to close by law or other state action. The definition of a trading day varies by region. We determined that the JSON format is the best suitable file format for the analysis dataset we created.

Data preprocessing can be defined as before the data mining models are operating, some fixes are applied on the dataset, fulfilling the missing data, clearing duplicate data, converting, combining, cleansing, normalizing, size reduction, etc. are transactions. We also preprocessed the tweet texts we collected in our study.

##### **Preprocessing steps**

- Lowercased texts.
- Removing all emojis.
- Removing all URLs.
- Removing all user tags with user info like @blabla.

- Removing all hashtags with hashtag info like #blabla.
- Removing all special characters %, \$, & etc.
- Removing all integers.
- Removing all strings whose length is less than 2.

In linguistic morphology and cognition, stemming is the process of reducing inflectional words to a stem, or root shape, usually a written word shape. We also passed all the tweets obtained after the preprocessing process through a stemmer for the consistency of our analysis. We used the Turkish Stemmer project for the stemming process, which is also compatible with the Turkish language (Otuncelli).

You can examine the result texts obtained because of the processes in Table 4.1. below.

Table 4.1 Some examples of Tweets before and after preprocessing

Original Tweet	Preprocessed and Stemmed Tweet
"Tasarrufun önemini biliyoruz. Enerji tasarrufu yaparak hem birikimlerimizi hem de doğamızı koruyabiliriz. 2020 yılında Genel Müdürlük binamız, bölge ve şubelerimizde yapmış olduğumuz çalışmalarla %14 oranında enerji tasarrufu sağladık. 🌱🌍 #EnerjiTasarrufuHaftası <a href="https://t.co/y8cgordrcA">https://t.co/y8cgordrcA</a> "	"tasarruf önem biliyor enerj tasarruf yaparak birikim doğa koruyabilir yıl genel müdürlük bina bölg şube yap olduk çalışma oran enerj tasarruf sağl"
"Yatırımlarınız Vadeli TL Mevduat Hesabı ile değer kazanıyor! Siz de yatırımlarınızı Vadeli TL Mevduat Hesabı'nda değerlendirin, cazip faiz oranları ile risk almadan kazanın! Detaylı bilgi için tıklayın."	"yatırım vade tl mevduat hesap il değer kazanıyor yatırım vade tl mevduat hesabı değerlendir cazip faiz oran il risk alma kaza detay bilgi iç tıklay"

Table 4.1 continuous

"Vadesiz hesabın rahatlığıyla, vadeli hesabın kazancını buluşturan Çift Sarılı Hesap, QNB Finansbank'ta! 🐱 Hemen başvurmak için: <a href="https://t.co/IAmJu8jaDq">https://t.co/IAmJu8jaDq</a> "	"vades hesap rahatlık vade hesap kazanç buluşturan çift sarı hesap qnb finansbank'ta hemen başvurmak iç"
--	--

The momentary changes in the inner world as a result of the events experienced or witnessed by the person are called sentiments. Sentiment analysis, which is also a part of text analysis, purposes to define the class (positive, negative, and neutral) that the dedicated text requests to state emotionally.

We also performed sentiment analysis on the processed texts we obtained within the scope of the study. We chose the Keyword Processing method for sentiment analysis. In Keyword Processing a sensitivity or belonging score is assigned to specific vocables or vocable sets according to their sentiment classes. These scores are used to calculate the total weight of the text/documents. It is the sum of the scores of the expressions in the total emotional weight of a text/document. Thus, with the final total score, it can be found to which class the text/document belongs. We used some example datasets from Kaggle (Tatman, 2017) for sentiment lexicons and applied the same preprocessing and stemming techniques to these positive and negative lexicon files. For the sentiment decision, we said that if positive keywords are more common than negative ones then the tweet has positive sentiment, if negative keywords are more common than positive ones then the tweet has negative sentiment and if both are equal then we said that the tweet sentiment is neutral.

You can find examples of tweets and the sentiment results in Table 4.2.

Table 4.2 Some examples of sentiments for Tweets

Tweet	Sentiment
<p>"Güvenli Ödeme Sistemi Albaraka'da! 2. el araç alıp satmanın en hızlı ve en güvenli yolu olan Albaraka Güvenli Ödeme Sistemi hakkında detaylı bilgi ve başvuru için linke tıklayabilirsiniz. <a href="https://t.co/fDiOfsdJwS">https://t.co/fDiOfsdJwS</a> <a href="https://t.co/edgRfKYzaa">https://t.co/edgRfKYzaa</a>"</p>	Positive
<p>"Milyonlarca kadının gördükleri şiddet karşısında susma sebepleri farklı, çünkü yaşları, statüleri, ekonomik durumları farklı. Ama gördükleri şiddet karşısında atmaları gereken adım her zaman aynı! #Susma #Saklama! #25KasımKadınaŞiddeteHayır #ŞiddetBulaşıcıdır <a href="https://t.co/bUy6nMGss0">https://t.co/bUy6nMGss0</a>"</p>	Negative
<p>"Enflasyon Korumalı Ara Dönem Ödemeli Mevduat Hesabı ile anaparanızı enflasyona karşı koruyun, kazanın. Detaylı bilgi için tıklayın."</p>	Neutral

We also calculated positive, negative, and neutral sentiment word counts for each tweet for further analysis.

Within the scope of our study, we also divided the daily changes in currency exchange and gold prices into 3 classes. We named these classes positive, negative, and neutral. By examining the movements in daily changes, we added the class labels that we determined on the data. In our dataset, we also calculated the 3-day historical change values. In this way, we had the opportunity to examine whether past change movements affected pricing.

After all these processes, our datasets to be used for analysis were prepared within the scope of 10 different banks and 3 different data groups. The final dataset example object is shown in Appendix 1.

## 4.2 Classification and Deep Learning

In classification, a dataset is assigned to a distinct and prearranged classes. The purpose of classification algorithms is to learn which data from the current training set will be assigned to which class. After training, tries to label test data correctly. We can also call the values indicating the data classes as labels.

We chose 7 different algorithms for classification within the scope of the study. These algorithms are Decision Tree, Multilayer Perceptron, Naive Bayes, Logistic Regression, Random Forest, Support Vector Machine, and K-Nearest Neighbor, respectively. For stock market movements, we labeled the changes under 3 different classes and identified these classes as positive, negative, and neutral.

Decision tree classification is a classification technique that creates a model in the shape of a tree format which contains decision nodes and leaf nodes by feature and target. The aim of this algorithm is separating the dataset into smaller parts (Uzun, 2019). We choose the Gini index as a quality measure and no pruning as a pruning method. We select binary nominal splits and a maximum number of nominals as 10.

A multilayer perceptron (MLP) is a feedforward artificial neural network that produces a set of outputs from a set of inputs. An MLP is described by a few layers of input nodes bonded as a directed graph between the input and output layers. MLP uses backpropagation to train the network (Science, 2022). We choose the maximum number of iterations count as 100, the number of hidden layers as 3, and the number of hidden neurons per layer as 10.

The Naive Bayes classifier is based on the Bayes theorem It is a lazy learning algorithm and can also run-on unstable datasets. Algorithm computes the possibility of every condition for a component and classifies it in accordance with the maximum possibility value. It is possible to create very successful studies with a small training data. Let's assume that a value in the test set has an unobservable value in the training set, it is possibility value equal to 0. As a result, it is not possible to estimate. This

situation is generally accepted as Zero Frequency. Corrective methods can be used to sort out this case. First correction method that comes to mind is known as Laplace estimation which is also simple than the others. (Hatipoglu, 2018). We choose default probability as 0.0001, minimum standard deviation as 0.0001 and maximum number of unique nominal values per attribute as 20.

Logistic Regression is a regression technique for classification. It is used to classify categorical or numerical data. It runs if the linked variable, namely the outcome, can only take 2 distinct values. We used logistic regression to define positive or negative in our study (Hatipoglu, 2018). We select stochastic average gradient as a solver.

Random Forest is a supervised learning algorithm. As the name suggests anyway, it generate a forest and does it anyway. The established "forest" is a set of decision trees, which are frequently trained by the "bagging" technique. According to the bagging technique, in generally combination of learning models raises the overall result. Briefly, a random forest creates more than one decision trees and consolidates them to get a more precise and balanced estimation (DevHunter, 2021). We choose the Gini index as a split criterion, tree depth to 10, and minimum node size 2. For forest options select 50 as the number of models and use a static random seed value.

Support Vector Machine is one of the supervised learning algorithms which is usually used in classification issues. It constructs a line to distinct spots placed on a plane and intends to have this line at the maximum space for the spots of desired classes. It can be used in small to medium complex datasets (Akca, 2020). We choose overlapping penalty as 1.0 and set RBF value to sigma 0.1.

In the simplest sense, KNN grounded predicting the class of the vector composed by the independent variables of the value to be predicted, based on the info in which class the nearest neighbors are frequent.

KNN (K-Nearest Neighbors) Algorithm does estimations on two basic variables;



- *Distance*: The distance of the point to be predicted from other points is calculated. For this, the Minkowski distance calculation function is used.
- *K (number of neighbors)*: We tell how many nearest neighbors to calculate. K value will immediately impress the outcome. If K is 1, the possibility of overfitting will be very high. It will give very general outcomes if K value is too large. For this reason, predicting the optimal K value remains the main subject of the problem (Arslan, 2020). We choose the number of neighbors to consider as 7 and weighted neighbors by distance.

For Deep Learning we defined a separate MLP with Tensorflow. We choose 50 hidden layers and ReLu as activation for the graph. For the output tensor, we used softmax as an activator. For the training phase, we choose batch sizes equal to 32 and 1000 epochs.

For all algorithms, we select 10-fold cross-validation and stratified sampling method.

We created 6 different models within the scope of our study. We determined these models by creating various variations using the values in the dataset we created. The model details created are shown in Table 4.3. below.

Table 4.3 Summary of created models with details

Model Name	Model Details
Model 1	Currency Exchange Prices and Gold Prices
Model 2	Model 1 + Stock Change Class 1, 2 and 3 days ago
Model 3	Model 1 + Sentiment Classes
Model 4	Model 3 + Normalized tweet reply counts, Normalized tweet retweet counts, Normalized tweet like counts
Model 5	Model 1 + Model 2 + Model 3
Model 6	Model 1 + Model 2 + Model 3 + Model 4

We classified our models with selected classifiers and collected results with 3 different categories and 10 different banks for 10 banks traded in BIST 100.



## CHAPTER 5

### RESULTS

In this section, we will analyze the data in the dataset we created according to the models we have determined. For the classification process, we created different models on KNIME and performed our analyzes on these models. Knime is an open-source and can be used for different types of computers and software packs (distributed to more than one business system) data analysis, reporting, and integration platform. Knime's components called "nodes" work with a drag and drop method. You can perform basic data preprocessing for visualization, modeling, and data analysis without writing code with Node structures.

In our proposed models, if we consider the first 3 most striking examples among the results obtained as a result of our analyses, among all tweets with the keyword “Halkbank”, Model 6 has achieved an accuracy rate of 69% with the deep learning algorithm as shown in Figure 5.1.

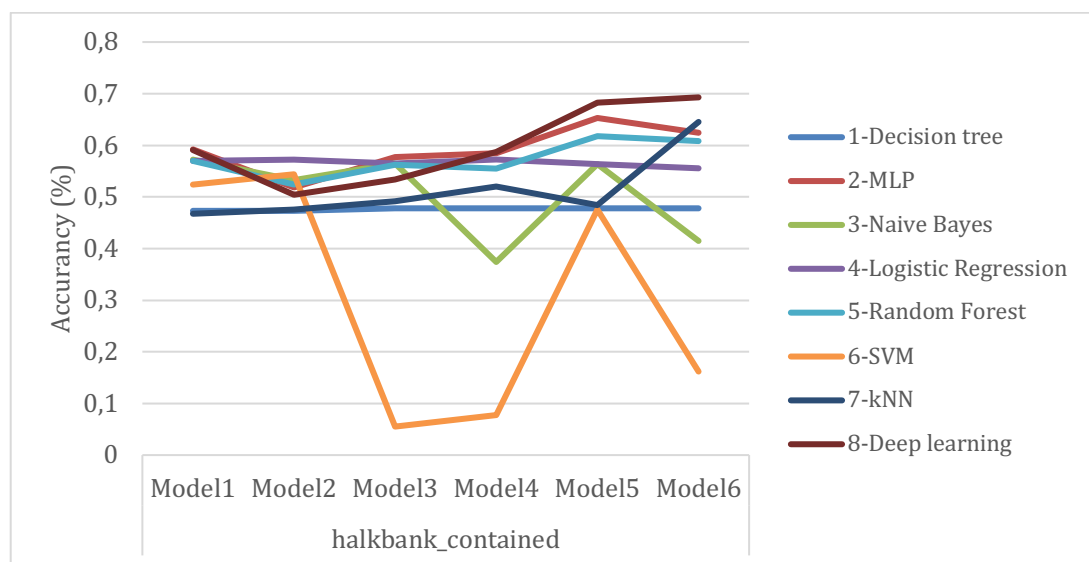


Figure 5.1 Accuracy of the Halkbank contained for the Model 6

Among all tweets with the keyword “GarantiBBVA”, Model 5 has achieved an accuracy rate of 68% with the deep learning algorithm as shown in Figure 5.2.

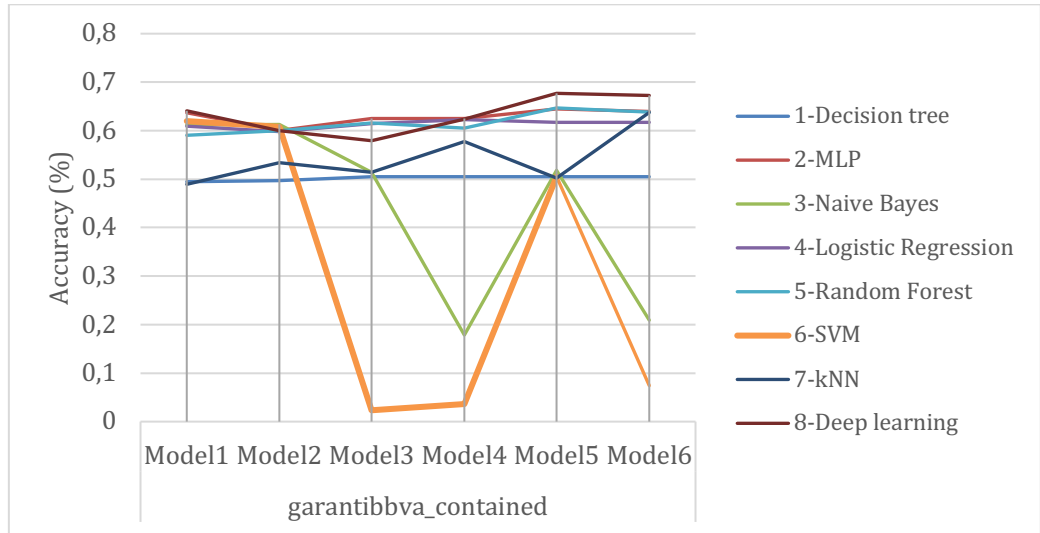


Figure 5.2 Accuracy of the GarantiBBVA contained for the Model 5

Among all tweets with the keyword “Vakıfbank”, Model 6 has achieved an accuracy rate of 66% with the deep learning algorithm as shown in Figure 5.3.

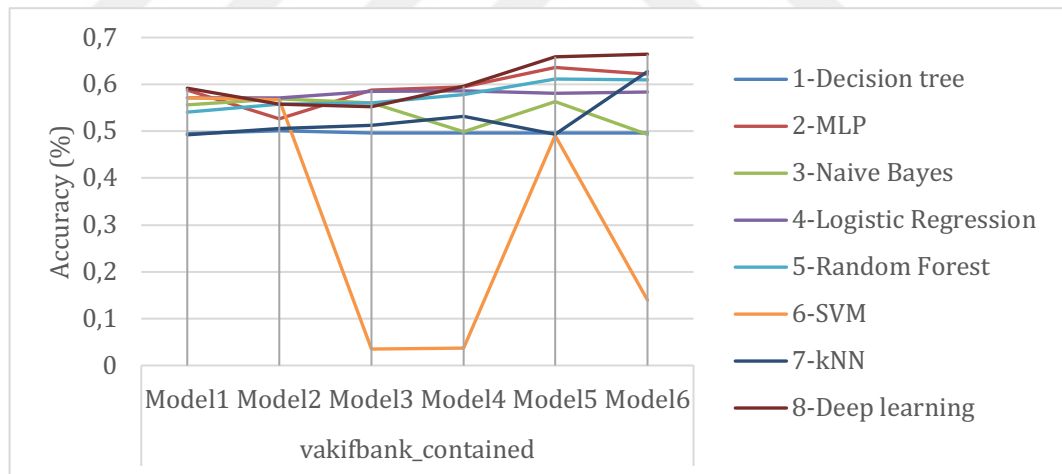


Figure 5.3 Accuracy of the Vakıfbank contained for the Model 6

As a result of our analysis, we have 8 different analysis methodologies, 3 different data sources, 6 different models, and 10 different banks which lead us to 1.4K distinct analysis results. We have listed the models and the algorithms that give the best results and their results under the subheadings of models.

## 5.1 Model 1

In Model 1, we examined the effect of daily currency exchange and daily gold prices on selected 10 banks' stock market prices.

According to the results we have obtained, the dataset containing all tweets that contain the keyword “GarantiBBVA” has obtained the highest accuracy value for Model 1 with 64%. MLP and Deep Learning were the classification methods that achieved the highest accuracy among the 30 different data sources we determined. In Model 1, the bank with the highest average accuracy %60 in stock market value estimation was determined as GarantiBBVA. Model 1 classification models and accuracies listed for used data sources in Table 5.1.

Table 5.1 Model 1 classification results

<b>Data Source</b>	<b>Classification Method</b>	<b>Accuracy</b>
akbank	Deep Learning	0.48
akbank_contained	Deep Learning	0.61
akbank_contained_verified	Deep Learning	0.57
albaraka	Random Forest	0.55
albaraka_contained	MLP	0.54
albaraka_contained_verified	SVM	0.59
garantibbva	Logistic Regression	0.53
garantibbva_contained	Deep Learning	0.64
garantibbva_contained_verified	Naive Bayes	0.62
halkbank	SVM	0.59
halkbank_contained	MLP	0.59
halkbank_contained_verified	MLP	0.57
isbank	Logistic Regression	0.58
isbank_contained	MLP	0.59

Table 5.1 continuous

isbank_contained_verified	Naive Bayes	0.59
qnbfinansbank	Logistic Regression	0.51
qnbfinansbank_contained	Decision Tree	0.52
qnbfinansbank_contained_verified	SVM	0.62
sekerbank	Decision Tree	0.45
sekerbank_contained	MLP	0.48
sekerbank_contained_verified	Logistic Regression	0.61
tskb	Deep Learning	0.57
tskb_contained	MLP	0.5
tskb_contained_verified	Deep Learning	0.49
vakifbank	MLP	0.46
vakifbank_contained	Deep Learning	0.59
vakifbank_contained_verified	SVM	0.59
yapikredi	Naive Bayes	0.59
yapikredi_contained	Deep Learning	0.58
yapikredi_contained_verified	Logistic Regression	0.5

## 5.2 Model 2

In Model 2, we examined the effect of daily currency exchange and daily gold prices and 3 days past stock change classes on selected 10 banks' stock market prices.

According to the results we have obtained, the dataset containing all tweets that contain the keyword “Sekerbank” posted by verified accounts has obtained the highest accuracy value for Model 2 with 62%. SVM was the classification method that achieved the highest accuracy among the 30 different data sources we determined. In Model 2, the bank with the highest average accuracy %58 in stock market value estimation was determined as Isbank. Model 2 classification models and accuracies listed for used data sources in Table 5.2.

Table 5.2 Model 2 classification results

<b>Data Source</b>	<b>Classification Method</b>	<b>Accuracy</b>
akbank	Deep Learning	0.51
akbank_contained	Naive Bayes	0.57
akbank_contained_verified	Logistic Regression	0.56
albaraka	Logistic Regression	0.52
albaraka_contained	Naive Bayes	0.53
albaraka_contained_verified	SVM	0.6
garantibbva	SVM	0.52
garantibbva_contained	Naive Bayes	0.61
garantibbva_contained_verified	Naive Bayes	0.6
halkbank	SVM	0.58
halkbank_contained	Logistic Regression	0.57
halkbank_contained_verified	MLP	0.55
isbank	SVM	0.59
isbank_contained	Naive Bayes	0.58
isbank_contained_verified	Naive Bayes	0.58
qnbfinansbank	kNN	0.45
qnbfinansbank_contained	Decision Tree	0.51
qnbfinansbank_contained_verified	SVM	0.61
sekerbank	Decision Tree	0.44
sekerbank_contained	Logistic Regression	0.47
sekerbank_contained_verified	Naive Bayes	0.62
tskb	Naive Bayes	0.56
tskb_contained	Random Forest	0.48
tskb_contained_verified	Random Forest	0.48
vakifbank	Deep Learning	0.47

Table 5.2 continuous

vakifbank_contained	Logistic Regression	0.57
vakifbank_contained_verified	SVM	0.59
yapikredi	SVM	0.59
yapikredi_contained	SVM	0.55
yapikredi_contained_verified	SVM	0.53

### 5.3 Model 3

In Model 3, we examined the effect of daily currency exchange and daily gold prices and the sentiment of the tweets on selected 10 banks' stock market prices.

According to the results we have obtained, the dataset containing all tweets that contain the keyword “GarantiBBVA” and also all tweets that contain the keyword “GarantiBBVA” and posted by verified accounts has obtained the highest accuracy value for Model 3 with 62%. Logistic Regression was the classification method that achieved the highest accuracy among the 30 different data sources we determined. In Model 3, the bank with the highest average accuracy %59 in stock market value estimation was determined as Isbank. Model 3 classification models and accuracies listed for used data sources in Table 5.3.

Table 5.3 Model 3 classification results

Data Source	Classification Method	Accuracy
akbank	Random Forest	0.46
akbank_contained	MLP	0.6
akbank_contained_verified	Logistic Regression	0.55
albaraka	Logistic Regression	0.54



Table 5.3 continuous

albaraka_contained	Naive Bayes	0.52
albaraka_contained_verified	Logistic Regression	0.58
garantibbva	Logistic Regression	0.53
garantibbva_contained	MLP	0.62
garantibbva_contained_verified	Random Forest	0.62
halkbank	Naive Bayes	0.57
halkbank_contained	Logistic Regression	0.57
halkbank_contained_verified	Logistic Regression	0.54
isbank	Logistic Regression	0.58
isbank_contained	MLP	0.58
isbank_contained_verified	Naive Bayes	0.57
qnbfinansbank	Random Forest	0.48
qnbfinansbank_contained	Naive Bayes	0.54
qnbfinansbank_contained_verified	MLP	0.61
sekerbank	Decision Tree	0.44
sekerbank_contained	Logistic Regression	0.48
sekerbank_contained_verified	Logistic Regression	0.6
tskb	Naive Bayes	0.55
tskb_contained	MLP	0.48
tskb_contained_verified	Decision Tree	0.45
vakifbank	Decision Tree	0.46

Table 5.3 continuous

vakifbank_contained	MLP	0.59
vakifbank_contained_verified	Random Forest	0.57
yapikredi	Random Forest	0.57
yapikredi_contained	Logistic Regression	0.56
yapikredi_contained_verified	Random Forest	0.52

#### 5.4 Model 4

In Model 4, we examined the effect of daily currency exchange and daily gold prices sentiment of the tweets, normalized tweet reply counts normalized tweet retweet counts and normalized tweet like counts on selected 10 banks' stock market prices.

According to the results we have obtained, the dataset containing all tweets that contain the keyword “GarantiBBVA” has obtained the highest accuracy value for Model 4 with 62%. Logistic Regression was the classification method that achieved the highest accuracy among the 30 different data sources we determined. In Model 3, the bank with the highest average accuracy %58 in stock market value estimation was determined as GarantiBBVA. Model 4 classification models and accuracies listed for used data sources in Table 5.4.

Table 5.4 Model 4 classification results

<b>Data Source</b>	<b>Classification Method</b>	<b>Accuracy</b>
akbank	Logistic Regression	0.48
akbank_contained	MLP	0.6
akbank_contained_verified	Logistic Regression	0.56
albaraka	Logistic Regression	0.5
albaraka_contained	Deep Learning	0.52

Table 5.4 continuous

albaraka_contained_verified	MLP	0.54
garantibbva	Logistic Regression	0.52
garantibbva_contained	MLP	0.62
garantibbva_contained_verified	Random Forest	0.61
halkbank	Decision Tree	0.53
halkbank_contained	Deep Learning	0.59
halkbank_contained_verified	Logistic Regression	0.56
isbank	Random Forest	0.57
isbank_contained	Deep Learning	0.58
isbank_contained_verified	Logistic Regression	0.56
qnbfinansbank	Naive Bayes	0.44
qnbfinansbank_contained	Logistic Regression	0.51
qnbfinansbank_contained_verified	Logistic Regression	0.58
sekerbank	Decision Tree	0.44
sekerbank_contained	MLP	0.49
sekerbank_contained_verified	Logistic Regression	0.6
tskb	Logistic Regression	0.55
tskb_contained	Deep Learning	0.49
tskb_contained_verified	Random Forest	0.48
vakifbank	Decision Tree	0.46
vakifbank_contained	Deep Learning	0.6
vakifbank_contained_verified	MLP	0.56
yapikredi	Logistic Regression	0.57
yapikredi_contained	Logistic Regression	0.56
yapikredi_contained_verified	kNN	0.51

## 5.5 Model 5

In Model 5, we examined the effect of daily currency exchange and daily gold prices, 3 days past stock change classes, and sentiment of the tweets on selected 10 banks' stock market prices.

According to the results we have obtained, the dataset containing all tweets that contain the keyword “GarantiBBVA” and also all tweets that contain the keyword “Halkbank” has obtained the highest accuracy value for Model 5 with 68%. Deep Learning was the classification method that achieved the highest accuracy among the 30 different data sources we determined. In Model 5, the bank with the highest average accuracy %62 in stock market value estimation was determined as Halkbank. Model 5 classification models and accuracies listed for used data sources in Table 5.5.

Table 5.5 Model 5 classification results

<b>Data Source</b>	<b>Classification Method</b>	<b>Accuracy</b>
akbank	Naive Bayes	0.47
akbank_contained	Deep Learning	0.65
akbank_contained_verified	Logistic Regression	0.55
albaraka	Naive Bayes	0.51
albaraka_contained	Deep Learning	0.54
albaraka_contained_verified	Naive Bayes	0.55
garantibbva	Decision Tree	0.49
garantibbva_contained	Deep Learning	0.68
garantibbva_contained_verified	Random Forest	0.6
halkbank	Logistic Regression	0.57

Table 5.5 continuous

halkbank_contained	Deep Learning	0.68
halkbank_contained_verified	Deep Learning	0.61
isbank	Logistic Regression	0.59
isbank_contained	Deep Learning	0.63
isbank_contained_verified	Naive Bayes	0.57

## 5.6 Model 6

In Model 6, we examined the effect of daily currency exchange and daily gold prices, 3 days past stock change classes, the sentiment of the tweets, normalized tweet reply counts, normalized tweet retweet counts, and normalized tweet like counts on selected 10 banks' stock market prices. Model 6 classification models and accuracies listed for used data sources in Table 5.6.

According to the results we have obtained, the dataset containing all tweets that contain the keyword “Halkbank” has obtained the highest accuracy value for Model 6 with 69%. Deep Learning was the classification method that achieved the highest accuracy among the 30 different data sources we determined. In Model 6, the bank with the highest average accuracy %61 in stock market value estimation was determined as Halkbank.

Table 5.6 Model 6 classification results

Data Source	Classification Method	Accuracy
akbank	Random Forest	0.46
akbank_contained	Deep Learning	0.64
akbank_contained_verified	Logistic Regression	0.56

Table 5.6 continuous

albaraka	Logistic Regression	0.52
albaraka_contained	kNN	0.55
albaraka_contained_verified	Random Forest	0.56
garantibbva	MLP	0.52
garantibbva_contained	Deep Learning	0.67
garantibbva_contained_verified	Random Forest	0.6
halkbank	SVM	0.53
halkbank_contained	Deep Learning	0.69
halkbank_contained_verified	Deep Learning	0.6
isbank	Logistic Regression	0.58
isbank_contained	kNN	0.64
isbank_contained_verified	Logistic Regression	0.57
qnbfinansbank	SVM	0.46
qnbfinansbank_contained	Deep Learning	0.57
qnbfinansbank_contained_verified	Logistic Regression	0.6
sekerbank	Logistic Regression	0.46
sekerbank_contained	Random Forest	0.5
sekerbank_contained_verified	Logistic Regression	0.59
tskb	Naive Bayes	0.55
tskb_contained	Deep Learning	0.66

Table 5.6 continuous

tskb_contained_verified	Random Forest	0.45
vakifbank	Random Forest	0.47
vakifbank_contained	Deep Learning	0.66
vakifbank_contained_verified	Deep Learning	0.59
yapikredi	Logistic Regression	0.57
yapikredi_contained	Deep Learning	0.61
yapikredi_contained_verified	MLP	0.53

## **CHAPTER 6**

### **CONCLUSION**

In this study, we tried to predict the stock market with the classification algorithms we have chosen with 6 different methods created by using Twitter data of 10 banks traded in Borsa Istanbul. For this reason, we collected social media, stock market and financial data in order to make relevant analysis. The collected data was passed through certain preprocessing steps and analyzed. The analysis models created were saved for future use.

According to the results obtained, GarantiBBVA is the most likely bank to make stock market predictions with Twitter data, followed by Halkbank. In addition, the accuracy rates of the tweets in which the bank's name is mentioned on all Twitter are higher than the other selected data groups. As a result of the study, algorithms with high validation rates can be listed as Deep Learning, MLP, Random Forest, and Logistic Regression.

According to all the findings obtained and all the models analyzed, there is a certain relationship between the currency exchange values of the country and the tweets posted on Twitter, and the stock market values of the selected banks.

In the future, the study can be diversified by increasing the selected data groups and creating new model options. Preprocessing steps can be changed and diversified. The parameters used in the analysis models can be retested under different conditions. We rely on that this study will be a good beginning spot for further studies and models.



## REFERENCES

- Akca, M. F. (2022, May 25). *Nedir Bu Destek Vektör Makineleri? (Makine öğrenmesi serisi-2)*. Medium. <https://medium.com/deep-learning-turkiye/nedir-bu-destek-vekt%C3%B6r-makineleri-makine-%C3%B6%C4%9Frenmesi-serisi-2-94e576e4223e>
- Akmese, H., Aras, S., & Akmese, K. (2016). Financial performance and social media: A research on tourism enterprises quoted in Istanbul stock exchange (BIST). *Procedia Economics and Finance*, 39, 705-710.
- Arslan, E. (2022, May 25). *Makine öğrenmesi-KNN (K-nearest neighbors) algoritması*. Medium. <https://medium.com/%40arslanev/makine-%C3%B6%C4%9Frenmesi-knn-k-nearest-neighbors-algoritmas%C4%B1-bdfb688d7c5f>
- Blankespoor, E., Miller, G. S., & White, H. D. (2014). The role of dissemination in market liquidity: Evidence from firms' use of Twitter™. *The Accounting Review*, 89(1), 79-112.
- Carosia, A. E. O., Coelho, G. P., & Silva, A. E. A. (2020). Analyzing the Brazilian financial market through Portuguese sentiment analysis in social media. *Applied Artificial Intelligence*, 34(1), 1-19.
- Chahine, S., & Malhotra, N. K. (2018). Impact of social media strategies on stock price: the case of Twitter. *European Journal of Marketing*, 52, 1526-1549.
- Chen, S., Gao, T., He, Y., & Jin, Y. (2019). Predicting the stock price movement by social media analysis. *Journal of Data Analysis and Information Processing*, 7(4), 295-305.
- Chen, W., Yeo, C. K., Lau, C. T., & Lee, B. S. (2018). Leveraging social media news to predict stock index movement using RNN-boost. *Data & Knowledge Engineering*, 118, 14-24.

Coyne, S., Madiraju, P., & Coelho, J. (2017, November). *Forecasting stock prices using social media analysis*. 2017 IEEE 15th Intl Conf on Dependable, Autonomic and Secure Computing, 15th Intl Conf on Pervasive Intelligence and Computing, 3rd Intl Conf on Big Data Intelligence and Computing and Cyber Science and Technology Congress (DASC/PiCom/DataCom/CyberSciTech) (pp. 1031-1038). IEEE.

DevHunter, Y. (2022, May 25). *Rastgele Orman(Random Forest) algoritması*. DEVHUNTER. <https://devhunteryz.wordpress.com/2018/09/20/rastgele-ormanrandom-forest-algoritmasi/>

Hatipoglu, E. (2022, May 25). *Machine learning - classification - logistic regression - part 8*. Medium. <https://medium.com/@ekrem.hatipoglu/machine-learning-classification-logistic-regression-part-8-b77d2a61aae1>

Hatipoglu, E. (2022, May 25). *Machine learning-classification-naive Bayes-part 11*. Medium. <https://medium.com/@ekrem.hatipoglu/machine-learning-classification-naive-bayes-part-11-4a10cd3452b4>

Icy Science. (2022, May 25). *Çok Katmanlı Algılayıcı (MLP) nedir?*. <https://tr.theastrologypage.com/multilayer-perceptron>

Jiao, P., Veiga, A., & Walther, A. (2020). Social media, news media and the stock market. *Journal of Economic Behavior & Organization*, 176, 63-90.

Kaushik, B., Hemani, H., & Ilavarasan, P. V. (2017). Social media usage vs. stock prices: an analysis of Indian firms. *Procedia Computer Science*, 122, 323-330.

Khan, W., Ghazanfar, M. A., Azam, M. A., Karami, A., Alyoubi, K. H., & Alfakeeh, A. S. (2022). Stock market prediction using machine learning classifiers and social media, news. *Journal of Ambient Intelligence and Humanized Computing*, 13(3), 3433-3456.

- Lee, L. F., Hutton, A. P., & Shu, S. (2015). The role of social media in the capital market: Evidence from consumer product recalls. *Journal of Accounting Research*, 53(2), 367-404.
- Liu, L., Wu, J., Li, P., & Li, Q. (2015). A social-media-based approach to predicting stock comovement. *Expert Systems with Applications*, 42(8), 3893-3901.
- Liu, P., Xia, X., & Li, A. (2018). Tweeting the financial market: Media effect in the era of Big Data. *Pacific-Basin Finance Journal*, 51, 267-290.
- López-Cabarcos, M. Á., Piñeiro-Chousa, J., & Pérez-Pico, A. M. (2017). The impact technical and non-technical investors have on the stock market: Evidence from the sentiment extracted from social networks. *Journal of Behavioral and Experimental Finance*, 15, 15-20.
- Nassirtoussi, A. K., Aghabozorgi, S., Wah, T. Y., & Ngo, D. C. L. (2014). Text mining for market prediction: A systematic review. *Expert Systems with Applications*, 41(16), 7653-7670.
- Nguyen, T. H., Shirai, K., & Velcin, J. (2015). Sentiment analysis on social media for stock movement prediction. *Expert Systems with Applications*, 42(24), 9603-9611.
- Otuncelli. (2022, May 25). *Otuncelli/Turkish-Stemmer-Python: Turkish language stemmer for python*. GitHub. <https://github.com/otuncelli/turkish-stemmer-python>
- Rosati, P., Deeney, P., Cummins, M., Van der Werff, L., & Lynn, T. (2019). Social media and stock price reaction to data breach announcements: Evidence from US listed companies. *Research in International Business and Finance*, 47, 458-469.
- Ruan, Y., Durresi, A., & Alfantoukh, L. (2018). Using Twitter trust network for stock market analysis. *Knowledge-Based Systems*, 145, 207-218.

- Siikanen, M., Baltakys, K., Kanninen, J., Vatrapi, R., Mukkamala, R., & Hussain, A. (2018). Facebook drives behavior of passive households in stock markets. *Finance Research Letters*, 27, 208-213.
- Sun, A., Lachanski, M., & Fabozzi, F. J. (2016). Trade the tweet: Social media text mining and sparse matrix factorization for stock market prediction. *International Review of Financial Analysis*, 48, 272-281.
- Tatman, R. (2022, May 25). *Sentiment Lexicons for 81 languages*. Kaggle. <https://www.kaggle.com/datasets/rtatman/sentiment-lexicons-for-81-languages>
- Uzun, E. (2022, May 25). *Decision Tree (Karar Ağacı): Id3 algoritması - classification (SINIFLAMA)*. [https://erdincuzun.com/makine\\_ogrenmesi/decision-tree-karar-agaci-id3-algoritmasi-classification-siniflama/](https://erdincuzun.com/makine_ogrenmesi/decision-tree-karar-agaci-id3-algoritmasi-classification-siniflama/)
- Van Dieijen, M., Borah, A., Tellis, G. J., & Franses, P. H. (2020). Big data analysis of volatility spillovers of brands across social media and stock markets. *Industrial Marketing Management*, 88, 465-484
- Wu, G. G. R., Hou, T. C. T., & Lin, J. L. (2019). Can economic news predict Taiwan stock market returns?. *Asia Pacific Management Review*, 24(1), 54-59.
- Zhang, T., Yuan, Y., & Wu, X. (2020). Is microblogging data reflected in stock market volatility? Evidence from Sina Weibo. *Finance Research Letters*, 32, 101173.

## APPENDICES

### Appendix 1. Example object of the final dataset

```
{
  "id":1349021927491432456,
  "conversation_id":"1349021927491432456",
  "created_at":"2021-01-12 18:54:25 +03",
  "date":"2021-01-12",
  "time":"18:54:25",
  "timezone":"+0300",
  "user_id":331711843,
  "username":"yapikredi",
  "name":"Yapı Kredi",
  "place":"",
  "tweet":"Code.YapıKredi ile \"Girişimcilik Sohbetleri\" serimizin bu ha
  "language":"tr",
  "mentions":[ ],
  "urls":[ ],
  "photos":[ ],
  "replies_count":4,
  "retweets_count":1,
  "likes_count":8,
  "hashtags":[ ],
  "cashtags":[ ],
  "Stock":{
    "updated_date":"2021-01-12",
    "price":3.25,
    "open":3.27,
    "high":3.29,
    "low":3.22,
    "vol":120590000.0,
    "change":0.31,
    "bank_index":10,
    "change_class":"Positive",
    "change_class1":"Negative",
    "change_class2":"Positive",
    "change_class3":"Positive"
  },
  "Dollar":{
    "open":7.4844,
    "selling":7.4568,
    "highest":7.5183,
    "lowest":7.419,
    "close":7.4568,
    "change":-0.37,
    "updated_date":"2021-01-12",
    "change_class":"Negative",
    "change_class1":"Positive",
    "change_class2":"Positive",
    "change_class3":"Negative"
  },
  "Euro":{ },
  "Gold":{ },
  "preprocessed_tweet":"code yapıkre il girişimcilik sohbet ser haf konu
  "sentiment":"Positive"
}
```